开源数据集

更多好资源请上 AI 研习社 点击下载：

http://www.gair.link/page/resources

这份开源数据集是由 skymind.ai 发布的，内容包括生物识别、自然图像以及深度学习图像等数据集，还算是比较全面。各位研友们，可以点击收藏查阅：

新增数据集

- 开源生物识别数据：http://openbiometrics.org/

- Google Audioset：扩展了 632 个音频分类样本，并从 YouTube 视频中提取了 2，084，320 个人类标记的 10 秒声音片段。
- 地址：https://research.google.com/audioset/

- Uber 2B trip data：首次展示 2 百万公里的出行数据。
- 地址：https://movement.uber.com/cities

- Yelp Open Dataset：Yelp 数据集是用于 NLP 的 Yelp 业务、评论和用户数据的子集。
- 地址：https://www.yelp.com/dataset

- Core50：用于连续目标识别的新数据集和基准。
- 地址：https://vlomonaco.github.io/core50/

- Kaggle 数据集：https://www.kaggle.com/datasets

- Data Portal：http://dataportals.org/

- Open Data Monitor：https://opendatamonitor.eu/

- Quandl Data Portal：https://www.quandl.com/

- Mut1ny 头部/面部分割数据集：http://www.mut1ny.com/face-headsegmentation-dataset

- Github 上的优秀公共数据集：https://www.kdnuggets.com/2015/04/awesome-public-datasets-github.html

- 头部 CT 扫描数据集：491 次扫描的 CQ500 数据集。
- 地址：http://headctstudy.qure.ai/

自然图像数据集

- MNIST：手写数字图像。最常用的可用性检查。格式 25x25、居中、黑白手写数字。这是一项简单的任务——仅某部分适用于 MNIST，不意味着它有效。
- 地址：http://yann.lecun.com/exdb/mnist/

- CIFAR10 / CIFAR100：32x32 彩色图像，10/100 类。虽然仍有趣却不再常用的可用性检查。
- 地址：http://www.cs.utoronto.ca/~kriz/cifar.html

- Caltech 101：101 类物体的图片。
- 地址：http://www.vision.caltech.edu/Image_Datasets/Caltech101/

- Caltech 256：256 类物体的图片。
- 地址：http://www.vision.caltech.edu/Image_Datasets/Caltech256/

- STL-10 数据集：用于开发无监督特征学习、深度学习、自学习算法的图像识别数据集。像修改过的 CIFAR-10。
- 地址：http://cs.stanford.edu/~acoates/stl10/

- The Street View House Numbers (SVHN)：Google 街景中的门牌号码。可以把它想象成复现的户外 MNIST。

- 地址：http://ufldl.stanford.edu/housenumbers/

- NORB：玩具摆件在各种照明和姿势下的双目图像。
- 地址：http://www.cs.nyu.edu/~ylclab/data/norb-v1.0/

- Pascal VOC：通用图像分割/分类——对于构建真实世界图像注释不是非常有用，但对基线很有用。
- 地址：http://pascallin.ecs.soton.ac.uk/challenges/VOC/

- Labelme：带注释图像的大型数据集。
- 地址：http://labelme.csail.mit.edu/Release3.0/browserTools/php/dataset.php

- ImageNet：新算法的客观图像数据集（de-facto image dataset）。许多图像 API 公司都有来自其 REST 接口的标签，这些标签近 1000 类;WordNet; ImageNet 的层次结构。
- 地址：http://image-net.org/

- LSUN：具有很多辅助任务的场景理解（房间布局估计，显著性预测（saliency prediction）等），有关联竞赛。（associated competition）。

- 地址：http://lsun.cs.princeton.edu/2016/

- MS COCO：通用图像理解/说明，有关联竞赛。
- 地址：http://mscoco.org/

- COIL 20：不同物体在 360 度旋转中以每个角度成像。
- 地址：http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

- COIL100：不同物体在 360 度旋转中以每个角度成像。
- 地址：http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

- Google 开源图像：有 900 万张图像的网址集合，这些图像通过知识共享（Creative Commons）被标注成 6000 多个类别。
- 地址：https://research.googleblog.com/2016/09/introducing-open-images-dataset.html

地理空间数据

- OpenStreetMap：免费提供整个星球的矢量数据。它包含（旧版）美国人口普查局的数据。
- 地址：http://wiki.openstreetmap.org/wiki/Planet.osm

- Landsat8：整个地球表面的卫星视角图，每隔几周更新一次。
- 地址：https://landsat.usgs.gov/landsat-8

- NEXRAD：美国大气层的多普勒雷达扫描图。
- 地址： https://www.ncdc.noaa.gov/data-access/radar-data/nexrad

人工数据集

- Arcade Universe：一个人工数据集生成器，图像包含街机游戏 sprite，如 tetris pentomino / tetromino。该生成器基于 O. Breleux 的 bugland 数据集生成器。
- 地址：https://github.com/caglar/Arcade-Universe

- 以 Baby AI School 为灵感的数据集集合。
- 地址：http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/BabyAISchool

- Baby AI Shapes Dataset：区分 3 种简单形状。

- 地址：

  http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/BabyAIShapesDatasets
- Baby AI Image And Question Dataset：一个问题-图像-答案数据集。
- 地址：

  http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/BabyAIImageAndQuestionDatasets

- Deep Vs Shallow Comparison ICML2007：为实证评估深层架构而生成的数据集。
- 地址：

  http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/DeepVsShallowComparisonICML2007

- MnistVariations：在 MNIST 中引入受控变化。
- 地址：

  http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/MnistVariations

- RectanglesData：区分宽矩形和垂直矩形。

- 地址：

  http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/RectanglesData

- ConvexNonConvex：区分凸形和非凸形状。
- 地址：

  http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/ConvexNonConvex

- BackgroundCorrelation：嘈杂 MNIST 背景下相关度的控制
- 地址：

  http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/BackgroundCorrelation

人脸数据集

- Labelled Faces in the Wild：13000 个经过裁剪的人脸区域（使用已经用名称标识符标记过的 Viola-Jones）。数据集中每个人员的子集里包含两个图像——人们常用此数据集训练面部匹配系统。
- 地址：http://vis-www.cs.umass.edu/lfw/

- UMD Faces：有 8501 个主题的 367，920 个面孔的带注释数据集。

- 地址：http://www.umdfaces.io/

- CASIA WebFace：超过 10，575 个人经面部检测的 453，453 张图像的面部数据集。需要一些质量过滤。
- 地址：http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html
- MS-Celeb-1M：100 万张全世界的名人图片。需要一些过滤才能在深层网络上获得最佳结果。
- 地址：https://www.microsoft.com/en-us/research/project/ms-celeb-1m-challenge-recognizing-one-million-celebrities-real-world/

- Olivetti：一些人类的不同图像。
- 地址：http://www.cs.nyu.edu/~roweis/data.html

- Multi-Pie：The CMU Multi-PIE Face 数据库。
- 地址：http://www.multipie.org/

- Face-in-Action：http://www.flintbox.com/public/project/5486/
- JACFEE：日本和白种人面部情绪表达的图像。
- 地址：http://www.humintell.com/jacfee/

- FERET：面部识别技术数据库。
- 地址：http://www.itl.nist.gov/iad/humanid/feret/feret_master.html

- mmifacedb：MMI 面部表情数据库。
- 地址：http://www.mmifacedb.com/

- IndianFaceDatabase：http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/

- 耶鲁人脸数据库 A：http://vision.ucsd.edu/content/yale-face-database
- 耶鲁人脸数据库 B：http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html

- Mut1ny 头部/面部分割数据集：像素超过 16K 的面部/头部分割图像
- 地址：http://www.mut1ny.com/face-headsegmentation-dataset

深度学习视频数据集

- Youtube-8M：用于视频理解研究的大型多样化标记视频数据集。
- 地址：https://research.googleblog.com/2016/09/announcing-youtube-8m-large-and-diverse.html

文本数据集

- 20 newsgroups：分类任务，将出现的单词映射到新闻组 ID。用于文本分类的经典数据集之一，通常可用作纯分类的基准或任何 IR /索引算法的验证。
- 地址：http://qwone.com/~jason/20Newsgroups/
- 路透社新闻数据集：（较旧）纯粹基于分类的数据集，包含来自新闻专线的文本。常用于教程。
- 地址：https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection

- 宾州树库：用于下一个单词或字符预测。
- 地址：http://www.cis.upenn.edu/~treebank/

- UCI's Spambase：来自著名的 UCI 机器学习库的（旧版）经典垃圾邮件数据集。根据数据集的组织细节，可以将它作为学习私人垃圾邮件过滤的基线。
- 地址：https://archive.ics.uci.edu/ml/datasets/Spambase

- Broadcast News：大型文本数据集，通常用于下一个单词预测。
- 地址：http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S44

- 文本分类数据集：来自 Zhang et al., 2015。用于文本分类的八个数据集合集。这些是用于新文本分类基线的基准。样本大小从 120K 至 3.6M 不等，范围从二进制到 14 个分类问题。数据集来自 DBPedia、亚马逊、Yelp、Yahoo！和 AG。
- 地址：https://drive.google.com/drive/u/0/folders/0Bz8a_Dbh9Qhbfll6bVpmNUtUcFdjYmF2SEpmZUZUcVNiMUw1TWN6RDV3a0JHT3kxLVhVR2M

- WikiText：来自维基百科高质量文章的大型语言建模语料库，由 Salesforce MetaMind 策划。
- 地址：http://metamind.io/research/the-wikitext-long-term-dependency-language-modeling-dataset/

- SQuAD：斯坦福问答数据集——应用广泛的问答和阅读理解数据集，其中每个问题的答案都以文本形式呈现。
- 地址：https://rajpurkar.github.io/SQuAD-explorer/

- Billion Words 数据集：一种大型通用语言建模数据集。通常用于训练分布式单词表征，如 word2vec。
- 地址：http://www.statmt.org/lm-benchmark/

- Common Crawl：网络的字节级抓取——最常用于学习单词嵌入。可从 Amazon S3 上免费获取。也可以用作网络数据集，因为它可在万维网进行抓取。
- 地址：http://commoncrawl.org/the-data/

- Google Books Ngrams：来自 Google book 的连续字符。当单词首次被广泛使用时，提供一种简单的方法来探索。
- 地址：https://aws.amazon.com/datasets/google-books-ngrams/

- Yelp 开源数据集：Yelp 数据集是用于 NLP 的 Yelp 业务、评论和用户数据的子集。
- 地址：https://www.yelp.com/dataset

深度学习文本问答数据集

- Maluuba News QA 数据集：CNN 新闻文章中的 12 万个问答对。
- 地址：https://datasets.maluuba.com/NewsQA

- Quora 问答对：Quora 发布的第一个数据集，包含重复/语义相似性标签。
- 地址：https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

- CMU Q / A 数据集：手动生成的仿真问/答对，维基百科文章对其难度评分很高。
- 地址：http://www.cs.cmu.edu/~ark/QA-data/

- Maluuba 面向目标的对话：程序性对话数据集，对话旨在完成任务或做出决定。常用于聊天机器人。
- 地址：https://datasets.maluuba.com/Frames

- bAbi：来自 Facebook AI Research（FAIR）的综合阅读理解和问答数据集。
- 地址：https://research.fb.com/projects/babi/

- The Children's Book Test：Project Gutenberg 提供的儿童图书中提取的（问题+背景、答案）对的基线。用于问答（阅读理解）和仿真查找。
- 地址：http://www.thespermwhale.com/jaseweston/babi/CBTest.tgz

深度学习情感数据集

- 多领域情绪分析数据集：较旧的学术数据集。
- 地址：http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

- IMDB：用于二元情感分类的较旧、较小数据集。对文献中的基准测试无法支持更大的数据集。
- 地址：http://ai.stanford.edu/~amaas/data/sentiment/

- Stanford Sentiment Treebank：标准情感数据集，在每个句子解析树的每个节点都有细粒度的情感注释。
- 地址：http://nlp.stanford.edu/sentiment/code.html

推荐和排名系统

- Movielens：来自 Movielens 网站的电影评分数据集，各类大小都有。
- 地址：https://grouplens.org/datasets/movielens/

- Million Song 数据集：Kaggle 上元数据丰富的大型开源数据集，可以帮助人们使用混合推荐系统。
- 地址：https://www.kaggle.com/c/msdchallenge

- Last.fm：音乐推荐数据集，可访问深层社交网络和其它可用于混合系统的元数据。
- 地址：http://grouplens.org/datasets/hetrec-2011/

- Book-Crossing 数据集：来自 Book-Crossing 社区。包含 278,858 位用户提供的约 271,379 本书的 1,149,780 个评分。
- 地址：http://www.informatik.uni-freiburg.de/~cziegler/BX/

- Jester：来自 73,421 名用户对 100 个笑话的 410 万个连续评分（分数从 -10 至 10）。
- 地址：http://www.ieor.berkeley.edu/~goldberg/jester-data/

- Netflix Prize：Netflix 发布了他们的电影评级数据集的匿名版；包含 480,000 名用户对 17,770 部电影的 1 亿个评分。首个主要的 Kaggle 风格数据挑战。随着隐私问题的出现，只能提供非正式版。
- 地址：http://www.netflixprize.com/

深度学习网络和图形

- Amazon Co-Purchasing：亚马逊评论从「购买此产品的用户也购买了……」这一部分抓取数据，以及亚马逊相关产品的评论数据。适合在网络中试行推荐系统。
- 地址：http://snap.stanford.edu/data/#amazon

- Friendster 社交网络数据集：在变成游戏网站之前，Friendster 以朋友列表的形式为 103,750,348 名用户发布了匿名数据。

- 地址：https://archive.org/details/friendster-dataset-201107

语音数据集

- 2000 HUB5 English：最近在 Deep Speech 论文中使用的英语语音数据，从百度获取。
- 地址：https://catalog.ldc.upenn.edu/LDC2002T43

- LibriSpeech：包含文本和语音的有声读物数据集。由多个朗读者阅读的近 500 小时的各种有声读物演讲内容组成，包含带有文本和语音的章节。
- 地址：http://www.openslr.org/12/

- VoxForge：带口音的清晰英语语音数据集。适用于提升不同口音或语调鲁棒性的案例。
- 地址：http://www.voxforge.org/

- TIMIT：英语语音识别数据集。
- 地址：https://catalog.ldc.upenn.edu/LDC93S1

- CHIME：嘈杂的语音识别挑战数据集。数据集包含真实、仿真和干净的录音。真实录音由 4 个扬声器在 4 个嘈杂位置的近 9000 个录音构成，仿真录音由多个语音环境和清晰的无噪声录音结合而成。

- 地址：http://spandh.dcs.shef.ac.uk/chime_challenge/data.html

- TED-LIUM：TED 演讲的音频转录。1495 个 TED 演讲录音以及这些录音的文字转录。
- 地址：http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus

深度学习音符音乐数据集

- Piano-midi.de: 古典钢琴曲
- 地址：http://www.piano-midi.de/

- Nottingham：超过 1000 首民谣
- 地址：http://abc.sourceforge.net/NMD/

- MuseData: 古典音乐评分的电子图书馆
- 地址：http://musedata.stanford.edu/

- JSB Chorales: 四部协奏曲
- 地址：http://www.jsbchorales.net/index.shtml

其它数据集

- CMU 动作抓取数据集：http://mocap.cs.cmu.edu/
- Brodatz dataset：纹理建模。

- 地址：http://www.ux.uis.no/~tranden/brodatz.html

- 来自欧洲核子研究中心的大型强子对撞机（LHC）的 300TB 高质量数据。
- 地址：
  http://opendata.cern.ch/search?ln=en&p=Run2011A+AND+collection:CMS-Primary-Datasets+OR+collection:CMS-Simulated-Datasets+OR+collection:CMS-Derived-Datasets

- 纽约出租车数据集：由 FOIA 请求而获得的纽约出租车数据，导致隐私问题。
- 地址：http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

- Uber FOIL 数据集：来自 Uber FOIL 请求的纽约 4.5M 拾取数据。
- 地址：https://github.com/fivethirtyeight/uber-tlc-foil-response

- Criteo 点击量数据集：来自欧盟重新定位的大型互联网广告数据集。
- 地址：http://research.criteo.com/outreach/

健康 &生物数据

- 欧盟传染病监测图集：http://ecdc.europa.eu/en/data-tools/atlas/Pages/atlas.aspx

- 默克分子活动挑战：http://www.kaggle.com/c/MerckActivity/data

- Musk dataset: Musk dataset 描述了以不同构造出现的分子。每个分子都是 musk 或 non-musk，且其中一个构造决定了这一特性。
- 地址：https://archive.ics.uci.edu/ml/datasets/Musk+(Version+2)

政府&统计数据

- Data USA: 最全面的可视化美国公共数据。
- 地址：http://datausa.io/

- 欧盟性别统计数据库：http://eige.europa.eu/gender-statistics

- 荷兰国家地质研究数据 ：http://www.nationaalgeoregister.nl/geonetwork/srv/dut/search#fast=i

ndex&from=1&to=50&any_OR_geokeyword_OR_title_OR_keyword=landinrichting*&relation=within

- 联合国开发计划署项目：http://open.undp.org/#2016